Global Assessment of Functioning

A Modified Scale

Sally Caldecott-Hazard, PH.D.
Richard C. W. Hall, M.D.

*The modified Global Assessment of Functioning (GAF) scale has more detailed criteria and a more structured scoring system than the original GAF. The two scales were compared for reliability and validity. Raters who had different training levels assigned hospital admission and discharge GAF scores from patient charts. Intraclass correlation coefficients for admission GAF scores were higher for raters who used the modified GAF (0.81) compared with raters who used the original GAF (0.62). Validity studies showed a high correlation (0.80) between the two sets of scores. The modified GAF also correlated well with Zung Depression scores *-0.73). The modified GAF may be particularly useful when interrater reliability needs to be maximum and/or when persons with varying skills and employment backgrounds – and without much GAF training – must rate patients. Because of the increased structures, the modified GAF may also be more resistant to rater bias. (Psychosomatics 1995; 36: _____-_____)*

Global severity of illness scales are important instruments for assessing change in psychiatric patients. Increasingly, such scales are being used [45] The scales are simple to administer and are more sensitive to differential treatment effects than measures of single dimensions of psychopathology. Probably the most often used global assessment instrument is the interviewer-rated Global Assessment of Functioning (GAF) scale, which is listed in the DSM-III-R as an Axis V diagnostic criterion test. This scale is very similar to Global Assessment Scale (GAS) developed by Endicott et al, [4] The Endicott scale has values that range from 1, representing the sickest patient, to 100, a person with no symptoms. The scale is divided into 10 equal intervals, with 10 scores in each interval, and the criteria that define each score in each interval are listed. The GAF scale has similar criteria and the same interval design, except that the value range is from 1 to 90 (absent or minimal symptoms), and there are 9 rather than 10 equal intervals.

by managed care companies and governmental agencies to determine who can and cannot be admitted to hospitals.

One potential problem is with both these interviewer-rated scales is that for the scores to be comparable and thus meaningful across different studies, interrater reliability in scoring must be quite consistent within a study and from one study to another. Interrater reliability is strongly influenced by two factors: 1) the consistency of the raters, and 2) the heterogeneity of patient illness severity. Endicott et al. tested the

[4] Received March 3, 1993; revised April 26, 1993; accepted May 21, 2993. From the Center for Psychiatry, Florida Hospital, Orlando, Florida, and Department of Psychiatry, University of Florida, Gainesville. Address reprint requests to Dr. Caldecott-hazard, Center for Psychiatry, Florida Hospital, 601 E. Rollings St., Orlando, FL 32803.

reliability of the GAS in 5 studies and reported intraclass correlation coefficients ranging from 0.61 to 0.91, with associated standard error of measurement scores ranging from 5.0 to 8.0 units. Most of Endicott's ratings were done by only a few, well-trained interviewers. Having consistently trained interviewers should produce a greater likelihood of higher interrated reliability scores and small standard errors. Yet even with this bias, two of the studies had intraclass correlation coefficients in the 0.60s, suggesting that the scale might be less reliable than had been hoped. In contrast, one of the reliability studies used 15 raters of different backgrounds and training levels. Although the intraclass correlation coefficient was high, this was due primarily to a greater heterogeneity of illness severity as compared to the other studies, not to interrater consistency of scoring. The lack of interrater consistency was demonstrated by a high standard error of measurement not seen in the other studies.

Although we could not find published reliability studies on the GAF in the literature, out subjective experience at Florida Hospital was that the GAF was used by staff members of different backgrounds (physicians with varying degrees of familiarity with the scale, nurses, Ph.D. researchers), and GAF rating from these staff differed substantially for the same patient. Thus, we hypothesized that the original GAF might be less reliable than we had expected. To test this hypothesis and to improve interrater reliability, we developed a modified GAF scale, and we formally tested interrater reliability in the original and modified versions of the GAF. We conducted our study in 1992-1993.

Methods

A modified GAF scale was developed by increasing the structure of the original GAF instrument with a greater number of criteria and with additional directions for assigning scores. We chose to modify the GAF rather than the GAS because the GAF, as listed in the DSM-III-R. reflects more current ideas on illness severity rating and is the more frequently used instrument. The criteria and scoring changes that we made in the GAF were tested among a small group of staff members who rated patients from successive drafts of the modified scale. When staff members had different rating of a given patient, their reasons were discussed and changes were made in the wording or use of the criteria or scoring directions.

Reliability studies for both the original and modified GAF scales were based on ratings of 16 patient intake histories and discharge summaries taken from the patients' hospital charts. All of these patients had diagnoses of major depression with or without comorbid eating disorders. They had all been inpatients on the Affective/Eating Disorders unit, and their intake histories were obtained by one of the same two doctors. These particular 16 patients were chosen for review because they had the most detailed intake histories and discharge summaries available. Thus, a maximum amount of patient information was available for evaluation with the GAF.

Two groups of staff from the psychiatric units at Florida Hospital rated each of the same 16 patient histories and discharge summaries. All patients were given a GAF score for the severity of illness at admission and a

second GAF score for illness severity at discharge. One group of staff rated the patients using the original GAF, and the other group of staff rated the patients using the modified GAF. None of the staff received any training in the use of either GAF, but they were allowed to read it and to ask questions for clarification. This procedure was followed to evaluate the consistency of ratings by untrained staff; therefore, we could evaluate the soundness and reliability of each GAF under these conditions.

The staff in the group using the original GAF consisted of 12 professionals (nurses, physicians, social workers, psychiatry technicians, and clinical Ph.D.'s) assigned to 2 inpatient treatment units (affective/eating disorders and psychiatric/medical). The staff in the group using the modified GAF consisted of another group of 12 professionals from other inpatient units (acute general psychiatry, adolescent, or intensive treatment). Within each of the rating groups (original or modified GAF), the means and standard errors were calculated for the ratings of each patient on admission and discharge. Intraclass correlation coefficients (ICC) were then calculated separately for the original GAF group on admission and discharge and for the modified GAF group on admission and discharge. Both the admission and discharge correlation coefficients were compared between the groups.

The concurrent validity of the modified GAF was tested by comparing admission scored of this instrument with admission scores of the instrument with admission scores on the original GAF, the Zung depression test, and a selfp0rating of global illness severity. Pearson Product Moment correlations were used for these assessments of validity. For the modified and original GAF comparison, admission scores were obtained from the same 16 patient histories and discharge summaries as in the reliability tests. For the modified GAF and Zung comparison and the modified GAF and self-rating of illness comparison, data were obtained from outpatient telephone interviews with 142 patients who and been discharged from Florida Hospital 6 months to 1.5 years before. These patients all had diagnoses of major depression with or without comoid diagnoses of eating disorder. Each patient had been evaluated using the modified GAF only, the Zung depression test, and a self-illness severity rating. The self-rated global illness scores were on a scale of 1-10, where 1 was sickest and 10 was most health.

RESULTS

The Modified GAF

The modified GAF retained the same 1-90 scale with the same 10-point intervals as the original GAF. All criteria in the original GAF were retained and were listed on separate lines to facilitate quick reading (Table 1). Additional criteria were added to most of the 10-point intervals and directions for scoring the patient's illness severity were added to most of the 10-point intervals, and directions for scoring the patient's illness severity were added at the end of each 10-point interval. The purpose of these additions was to decrease the variability in scoring. Usually, the scoring within a 10-point interval applied only to the criteria within that interval. For example, in the 81-90 interval, a patient having no

symptoms or problems received a score of 88-90; a patient having minimal symptoms or problems received a score of 84-87; and a patient having minimal symptoms and problems rece3ived a score of 81-83 (table 1). However, in the 21-30, 31-40, and 41-50 scoring intervals, the same 10 criteria were listed in each interval, and the score depended on the number of criteria that a patient met within these 3 scoring intervals

For example, if a patient met 1 of these criteria, the score was 48-50; if a patient met 2 of the criteria, the score was 44-47; and if the patient met 3 of the criteria, the score was 41-43. However, if the patient met 4-6 of the criteria, the scores ranged from 31-40. If the patient met 7-10 of the criteria, the scores ranged from 21-30 (Table 1). Finally, in the 21-30 scoring interval, a unique set o f criteria and scores also existed in addition to the criteria and scoring already discussed. These unique criteria were listed in the original GAF and were deemed to be of sufficient seriousness that they should not be added to the list of criteria in the 31-40 and 41-50 intervals but rather would warrant the lowest score available in the 21-30 category. Thus, suicidal preoccupation and preparation, behavior considerable influenced by delusions or hallucinations, or serious impairment in communications (i.e., sometimes incoherent or profound stuporous depression), always elicited a score of 21.

The various changes we made in modifying the GAF made it longer than the original GAF (4 pages vs. 1). Thus, it is suggested that when using this new GAF, the interviewer should question the patient about each of the criteria, then write down answers, and later count the number of criteria that the patient meets. It is felt that the slower speed in assigning a score from the modified GAF is compensated for by the increased consistency of ratings attributable.

Interesting, all of the means for the patient's admission GAF scores were also higher in the original GAF group than in the modified group. Thus, the modified GAF caused patients to be rated more sick than the original GAF.

Concurrent Validity

Because all of the mean admission GAF scores for the original group were higher than the scores in the modified group, we wanted to test the correlation between the scores of the two GAF's and test the correlation of the modified GAF with other psychological assessment tests. The Pearson Product Moment correlation coefficient between the 16 original and 16 modified mean admission scores was 0.80.p<0.0001,df-14, showing good correlation (Table 2).

Because all of the patients used in these studies were depressed, we also compared modified GAF scores with the scores from the Zung depression test. The Pearson Product Moment correlation coefficient was -0.73, l',0.001 (negative because a higher number represents sickness in the Zung scores and a lower number represents sickness in the GAF) (Table 2).

Finally, we also correlated modified GAF scores with the scores that patients gave themselves to indicate their severity of illness. The Pearson Product Moment correlation coefficient was 0.58, P,0.01 (Table 2).

DISCUSSION

Our findings of an intraclass correlation coefficient of 0.62 for admission scores on the original GAF agreed with Endicott et al.'s report of ICC's ranging from 0.61 to 0.91. Our ICC of 0.62 was significant at P,0.001, thus indicating that while the reliability was somewhat low for admissions ratings, it still was perfectly usable. Likewise, the ICC for discharge ratings from the original GAF was 0.90, which indicates excellent reliability. The value of the modified GAF (with its admission ICC of 0.81 and discharge ICC of 0l95) is for instances when interrater reliability needs to be as high as it can be or when multiple persons of varying employment backgrounds and without much GAF training will rate patients. Research is a prime example for both uses of the modified GAF. Usually during research studies, where would also be enough time to read this longer GAF and assign ratings.

Another use for the modified GAF, compared with the original GAF or GAS, is in evaluating the need for hospital admission. Specifically, Thompson et al., in a review of 9,055 adult intakes, found marked variations in the way managed care case managers, compared with providers, assigned GAS scores generated from the same data. Thompson and colleagues felt that higher (less sick) scores reflected a need by managed care companies to limit the use of all inpatient services rather than their desire to selectively eliminated unnecessary hospitalizations. The ability of the managed care industry to affect the GAS scores in this way is attributed to the relatively less-structured nature of the GAS instrument, leading to lower interrater reliability. As we have shown, the modified GAF is both more structured than the original GAF or GAS and has better interrater reliability on admission scores. Thus, the modified GAF is less likely to reflect a bias by a managed care or governmental agency.

In addition to reliability tests, modified GAF ratings were also correlated with Zung depression tests and self-rating of illness severity in outpatients. Similar to reliability tests, these correlations were in the same range as the correlations that Endicott et al. found between the original GAS and the Mental Status Examination Record (FEF) in outpatients. The slightly higher correlation between the modified GAF and the Zung depression test (-0.71), compared with the original GAS and MSER (0.62), probably was because all of out patients were depressed and the Zung specifically assessed depression. In contrast, the MSER is a global rating scale like the GAS, and there was probably greater heterogeneity among these patients. However, both of these sets of correlations were acceptable, thereby indicating that the interviewer rated scales provide similar types of information and the original GAS and modified GAF each show acceptable validity. Interestingly, both the self-rated illness severity test that we correlated with the modified GAF and the FEF, correlated by Endicott with the original GAS, gave scored based on someone other than the interviewer's judgment, specifically the patient or the patient's family. Both of these sets of correlations were fairly low, 0.58 for the self-rated scale and modified GAF and -0.52 or -0.45 for the FEF and the original GAS. While the Zung is also a self-rated instrument, it's questions are more objective than the self-rated global illness scale or FEF, which may have accounted for the Zung's higher correlation with the GAF. Still,

McGlashan and Pfeiffer reported that patient self-assessment and physician or interviewer assessments of patients may differ significantly. One might also expect the same discrepancy between family and interviewer assessment of patients. Thus, the interviewer vs. self or family-rating procedures for measuring severity of illness often cannot be considered as providing similar or redundant information.

The modified GAF is an instrument having a higher reliability and similar validity to the original GAF or GAS. The modified GAF may be particularly useful when interrater reliability needs to be maximum (i.e., in research or as a tool to determine need for hospitalization) and/or when multiple persons of varying skills and employment backgrounds and without having had much GAF training (i.e. in managed care organizations) must rate patients. In addition, when used to evaluate the need for hospital admission, the modified GAF is less likely than the original GAF or GAS to reflect a provider or managed care bias. Thus, our modified GAF may be a better and improved patient assessment tool, one that can more accurately reflect a patient's true need for hospitalization.